Standard errors and the bootstrap

Reference: *Data Science* Chapter 5

Outline

- Uncertainty quantification as a "what if?"
- Sampling distributions
- Bias and standard error
- The bootstrap
- Bootstrapped confidence intervals
- Bootstrapped versus plug-in standard errors
- Bonus topic: the parametric bootstrap

Quantifying uncertainty

From the New England Journal of Medicine in 2006:

> We randomly assigned patients with resectable adenocarcinoma of the stomach, esophagogastric junction, or lower esophagus to either perioperative chemotherapy and surgery (250 patients) or surgery alone (253 patients)….
>
> With a median follow-up of four years, 149 patients in the perioperative-chemotherapy group and 170 in the surgery group had died. As compared with the surgery group, the perioperative-chemotherapy group had a higher likelihood of overall survival (five-year survival rate, 36 percent vs. 23 percent).

Quantifying uncertainty

Conclusion:

- Chemotherapy patients are **13%** more likely to survive past 5 years.

Quantifying uncertainty

Conclusion:

- Chemotherapy patients are **13%** more likely to survive past 5 years.

Not so fast! In statistics, we ask "what if?" a lot:

- What if the randomization of patients just happened, by chance, to assign more of the healthier patients to the chemo group?
- Or what if the physicians running the trial had enrolled a different sample of patients from the same clinical population?

Quantifying uncertainty

Conclusion:

- Chemotherapy patients are **13%** more likely to survive past 5 years.

Always remember two basic facts about samples:

- *All numbers are wrong*: any quantity derived from a sample is just a *guess* of the corresponding population-level quantity.
- *A guess is useless without an error bar*: an estimate of how wrong we expect the guess to be.

Quantifying uncertainty

Conclusion:

- Chemotherapy patients are **13% ± ?** more likely to survive past 5 years, with **??%** confidence.

By "quantifying uncertainty," we mean filling in the blanks.

Quantifying uncertainty

In stats, we equate trustworthiness with *stability*:

- If our data had been different merely due to chance, would our answer have been different, too?
- Or would the answer have been stable, even with different data?

Confidence in your estimates $\iff$ Stability of those estimates under the influence of chance

Quantifying uncertainty

For example:

- If doctors had taken a different sample of 503 cancer patients and gotten a drastically different estimate of the new treatment's effect, then the original estimate isn't very trustworthy.
- If, on the other hand, pretty much any sample of 503 patients would have led to the same estimates, then their answer for *this particular* subset of 503 is probably accurate.

Let's work through a thought experiment…

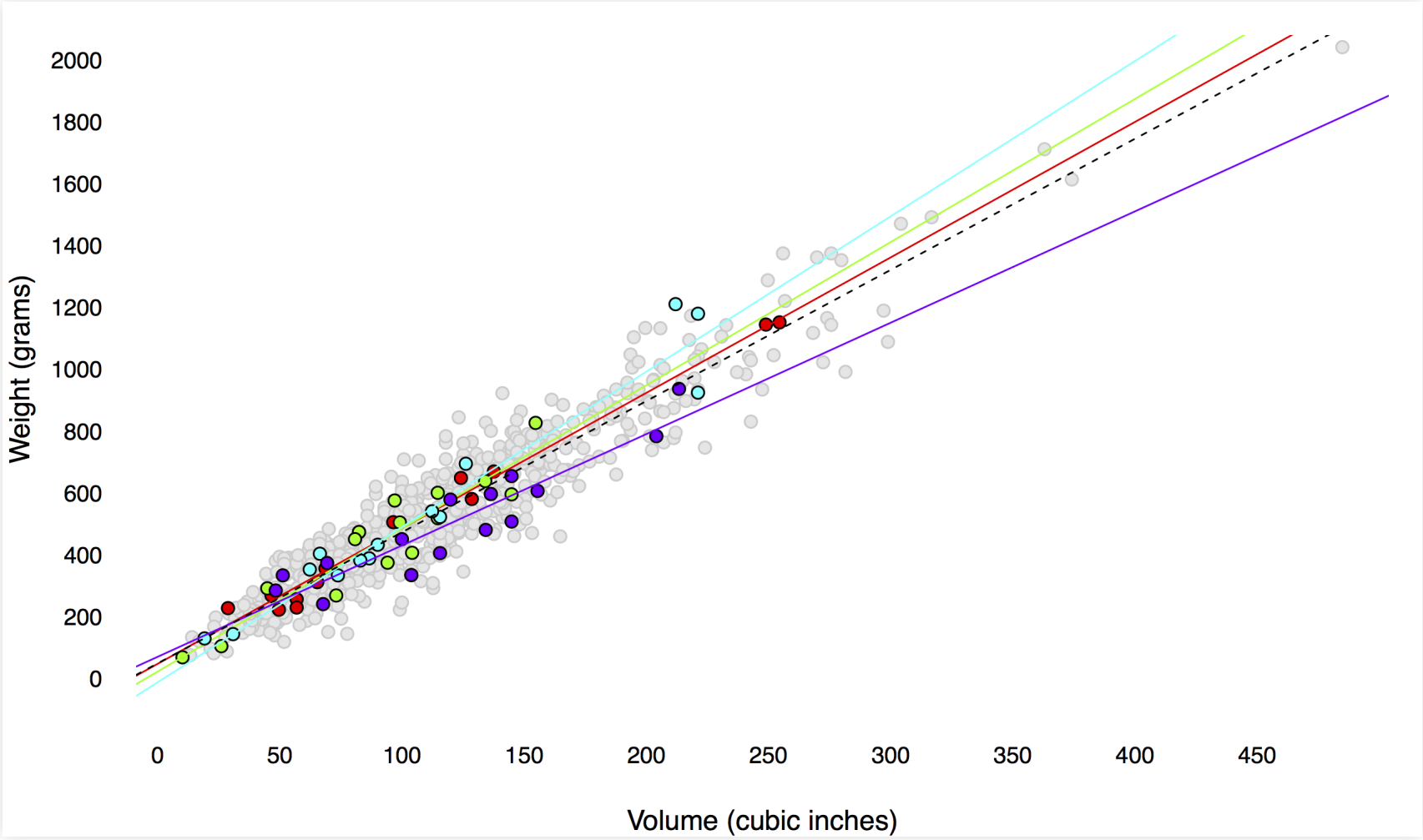
Kolmorogov goes fishing...

Kolmorogov goes fishing...

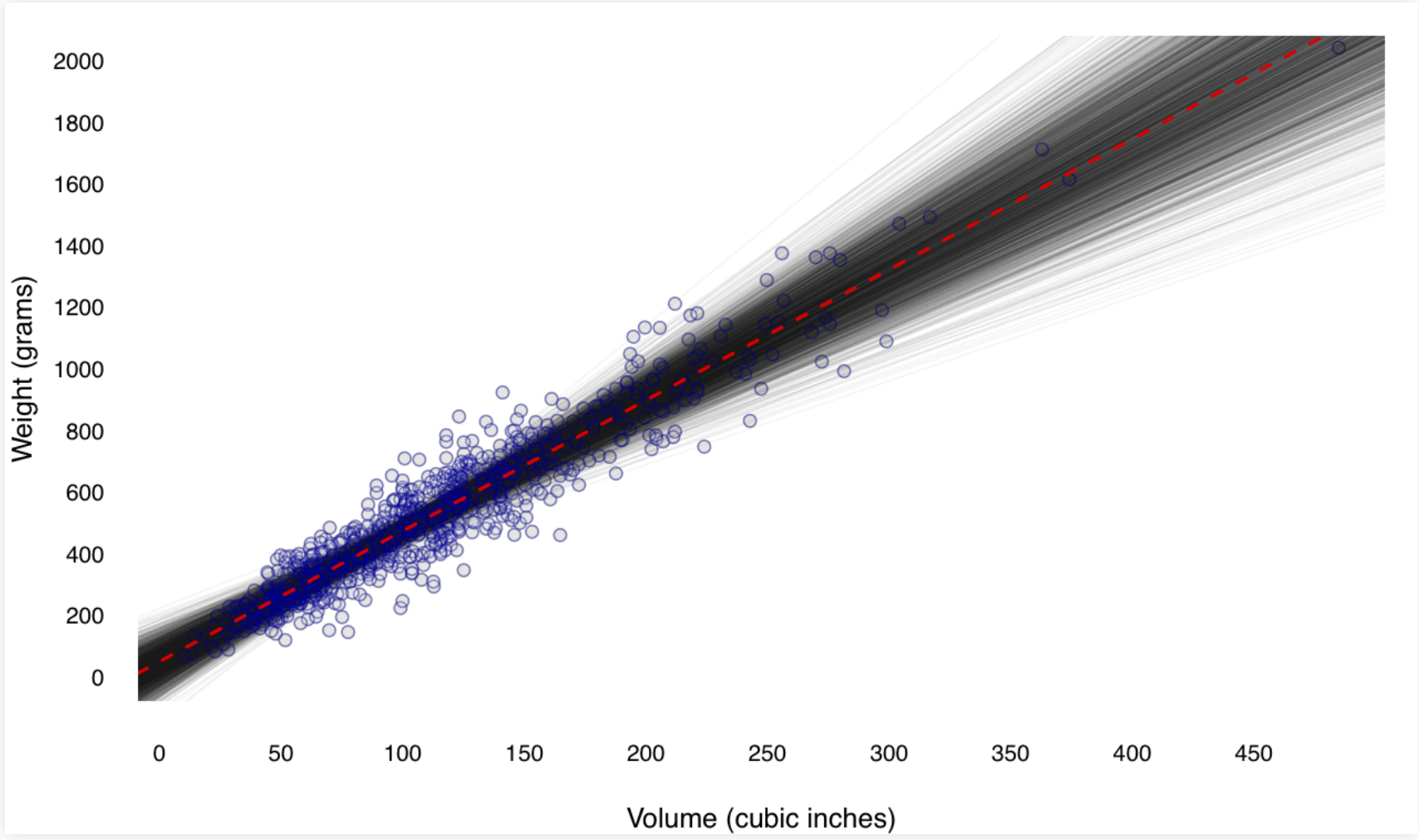
Imagine Andrey Kolmorogov on four-day fishing trip.

- The lake is home to a very large population of fish of varying size and weight.
- On each day, Kolmorogov takes a random sample of size $N = 15$ from this population—that is, he catches (and releases) 15 fish.
- He records the weight and approximate volume of each fish.
- He uses each day's catch to compute a different estimate of the volume–weight relationship for **all** fish in the lake.

Kolmorogov goes fishing...



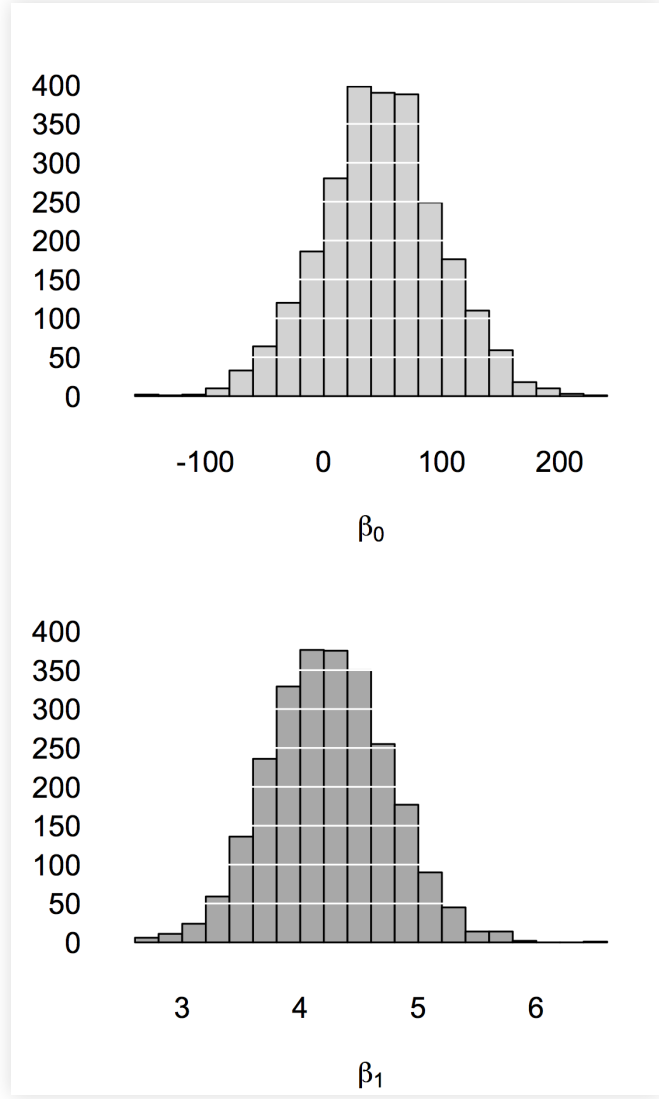
Kolmorogov goes fishing...



Kolmorogov goes fishing...

At right we see the *sampling distribution* for both β_0 and β_1.

- Each is centered on the true population value.
- The spread of each histogram tells us how *variable* our estimates are from one sample to the next.



Some notation

Suppose we are trying to estimate some population-level quantity θ: the *parameter* of interest.

So we take a sample from the population: $X_1, X_2, \ldots, X_N$.

We use the data to form an estimate $\hat{\theta}_N$ of the parameter. Key insight: $\hat{\theta}_N$ is a random variable.

Some notation

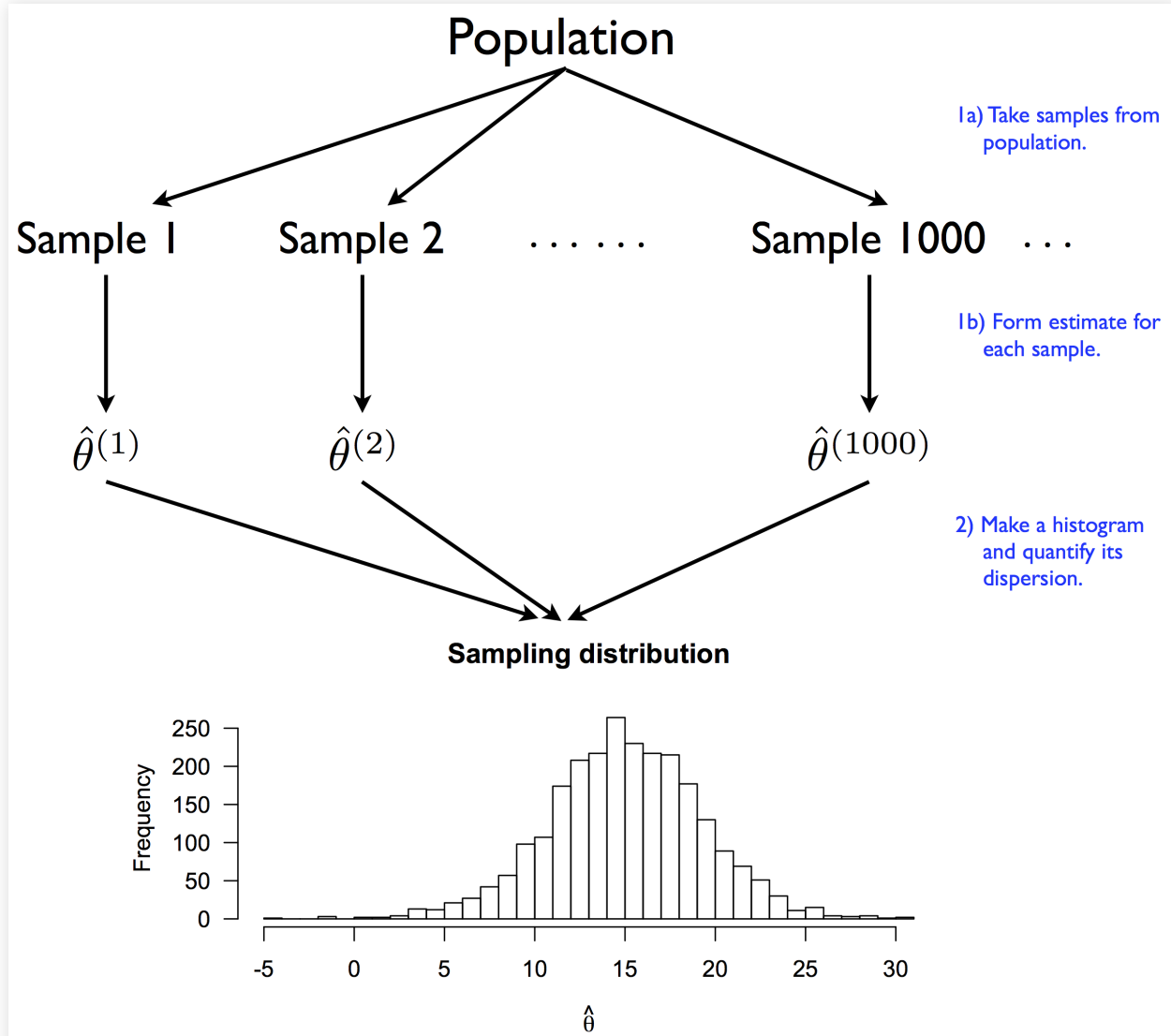
Suppose we are trying to estimate some population-level quantity θ: the *parameter* of interest.

So we take a sample from the population: $X_1, X_2, \ldots, X_N$.

We use the data to form an estimate $\hat{\theta}_N$ of the parameter. Key insight: $\hat{\theta}_N$ is a random variable.

Now imagine repeating this process thousands of times! Since $\hat{\theta}_N$ is a random variable, it has a probability distribution.

Some notation



Key definitions

Estimator: any method for estimating the value of a parameter (e.g. sample mean, sample proportion, slope of OLS line, etc).

Sampling distribution: the probability distribution of an estimator $\hat{\theta}_N$ under repeated samples of size N.

Bias: Let $\bar{\theta}_N = E(\hat{\theta}_N)$ be the mean of the sampling distribution. The bias of $\hat{\theta}_N$ is $(\bar{\theta}_N - \theta)$: the difference between the average answer and the truth.

Unbiased estimator: $(\bar{\theta}_N - \theta) = 0$.

Standard error

Standard error: the standard deviation of an estimator's sampling distribution:

$$\begin{aligned}\text{se}(\hat{\theta}_N) &= \sqrt{\text{var}(\hat{\theta}_N)} \\ &= \sqrt{E[(\hat{\theta}_N - \bar{\theta}_N)^2]} \\ &= \text{Typical deviation of } \hat{\theta}_N \text{ from its average}\end{aligned}$$

"If I were to take repeated samples from the population and use this estimator for every sample, how much does the answer vary, on average?"

Standard error

If an estimator is unbiased, then $\bar{\theta}_N = \theta$, so

$$\begin{aligned}
\text{se}(\hat{\theta}_N) &= \sqrt{E[(\hat{\theta}_N - \bar{\theta}_N)^2]} \\
&= \sqrt{E[(\hat{\theta}_N - \theta)^2]} \\
&= \text{Typical deviation of } \hat{\theta}_N \text{ from the truth}
\end{aligned}$$

"If I were to take repeated samples from the population and use this estimator for every sample, how big of an error do I make, on average?"

An analogy

THE WALL STREET JOURNAL.

HOMES

Farmhouse Fever Sweeps City Homes

Forget minimalism. Inspired by Chip and Joanna Gaines' 'Fixer Upper,' the urban farmhouse aesthetic is now in vogue; homes in such rustic styles can push up list prices as much as 30%

An analogy

This is why doctors and lawyers are buying “farmhouses.”

An analogy

Chip and Joanna lifestyle item #1: the farmhouse sink

The farmhouse idyll...

And the fine print

FEATURES

- Acid and alkali resistant glazed surfaces
- Suitable for waste disposal units or basket strainer waste
- Standard 3 1/2" diameter US drain opening (centered)
- Sink measures 30" x 18"
- Due to ±2% dimensions, it is recommended that the cabinet maker wait until your sink is delivered to make the cabinet because no template is available
- Includes “Shaws” blue badge
- Weight 155 lbs.
- 33" minimum cabinet size
- Hygienic due to antibacterial properties

Bowl options	Single bowl
Hole Configuration	1
Mounting Location	Deck
Installation Type	Apron front
Warranty	1-year limited warranty, 10-year limited warranty on fading/staining

Manufacturing tolerances

- On average across many weeks of manufacturing, the fancy sink has width equal to 30".
- But individual sinks vary from the average by about 0.5", due to manufacturing variability.
- So I expect that my specific sink will be somewhere in the vicinity of 30" $\pm$ 0.5".

Don't make any lifestyle choices that require greater precision!

Standard errors

- On average across many samples, my estimator $\hat{\theta}_N$ is equal to the right answer (θ).
- But individual estimates vary from the average by about $\mathrm{se}(\hat{\theta}_N)$, due to sampling variability.
- So I expect that the right answer is somewhere in the vicinity of $\theta_N \pm \mathrm{se}(\theta_N)$.

Don't reach any scientific conclusions that require greater precision!

Standard errors

But there's a problem here…

- Knowing the standard error requires knowing what happens across many separate samples.
- But we've only got our one sample!
- So how can we ever calculate the standard error?

Standard errors

> Two roads diverged in a yellow wood
> And sorry I could not travel both
> And be one traveler, long I stood
> And looked down one as far as I could
> To where it bent in the undergrowth…
>
> –Robert Frost, *The Road Not Taken*, 1916

Quantifying our uncertainty would seem to require knowing all the roads not taken—an impossible task.

The bootstrap

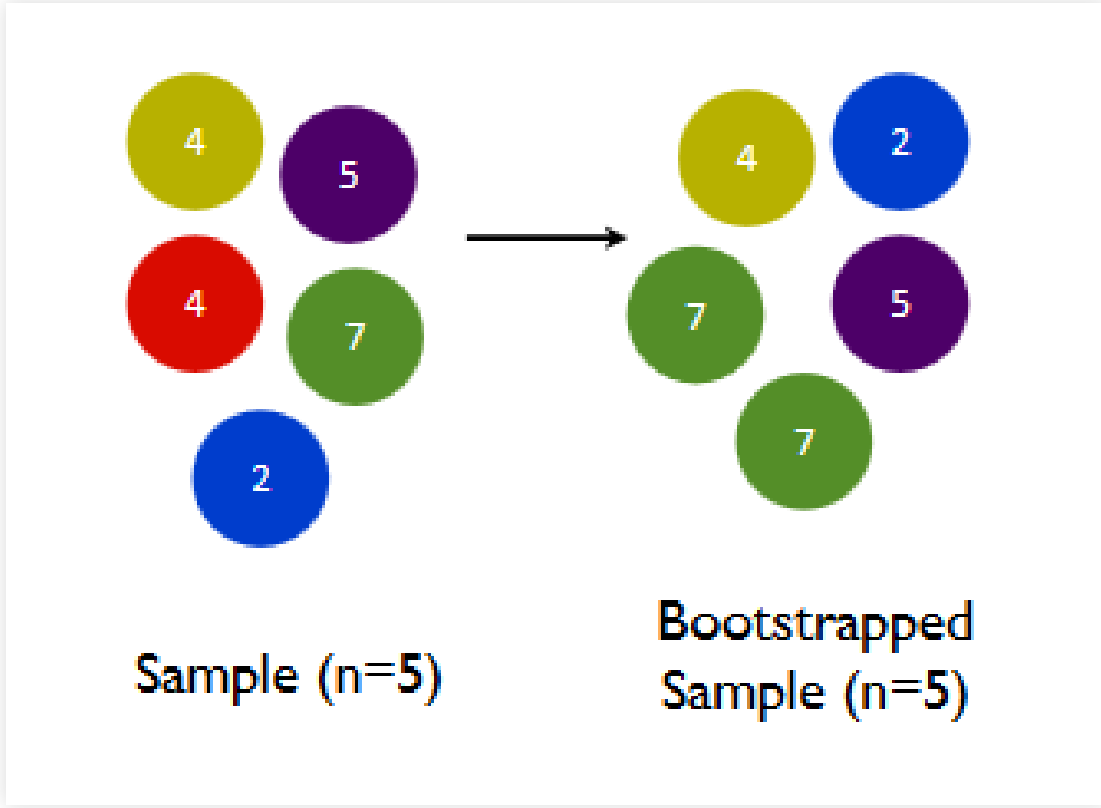
Problem: we can't take repeated samples of size N from the population, to see how our estimate changes across samples.

Seemingly hacky solution: take repeated samples of size N, with replacement, *from the sample itself*, and see how our estimate changes across samples. This is something we can easily simulate on a computer.

Basically, we pretend that our sample is the whole population and we charge ahead! This is called *bootstrap resampling*, or just *bootstrapping*.

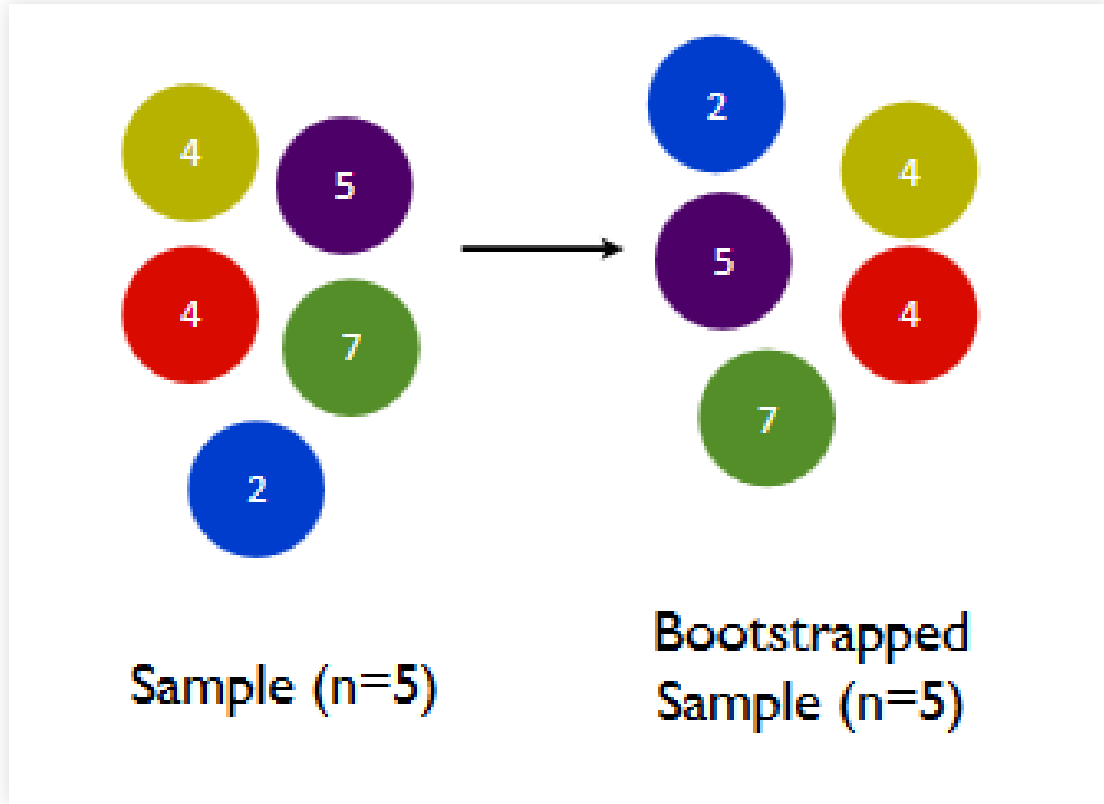
Sampling with replacement is key!

Bootstrapped sample 1



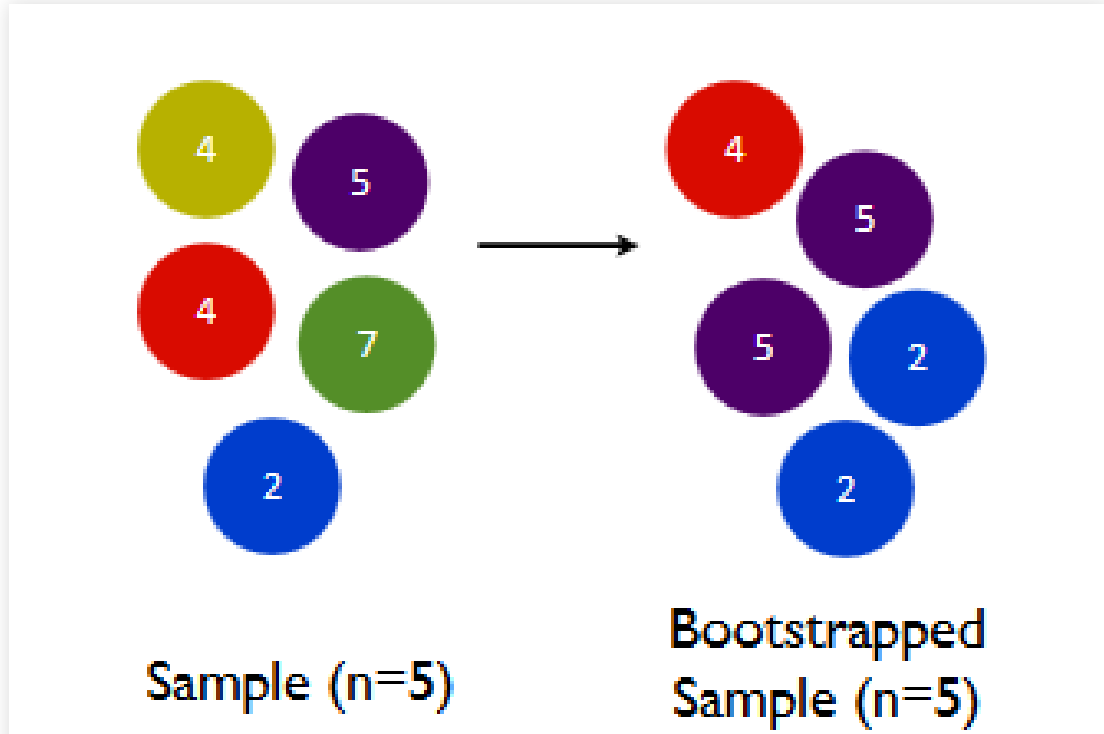
Sampling with replacement is key!

Bootstrapped sample 2

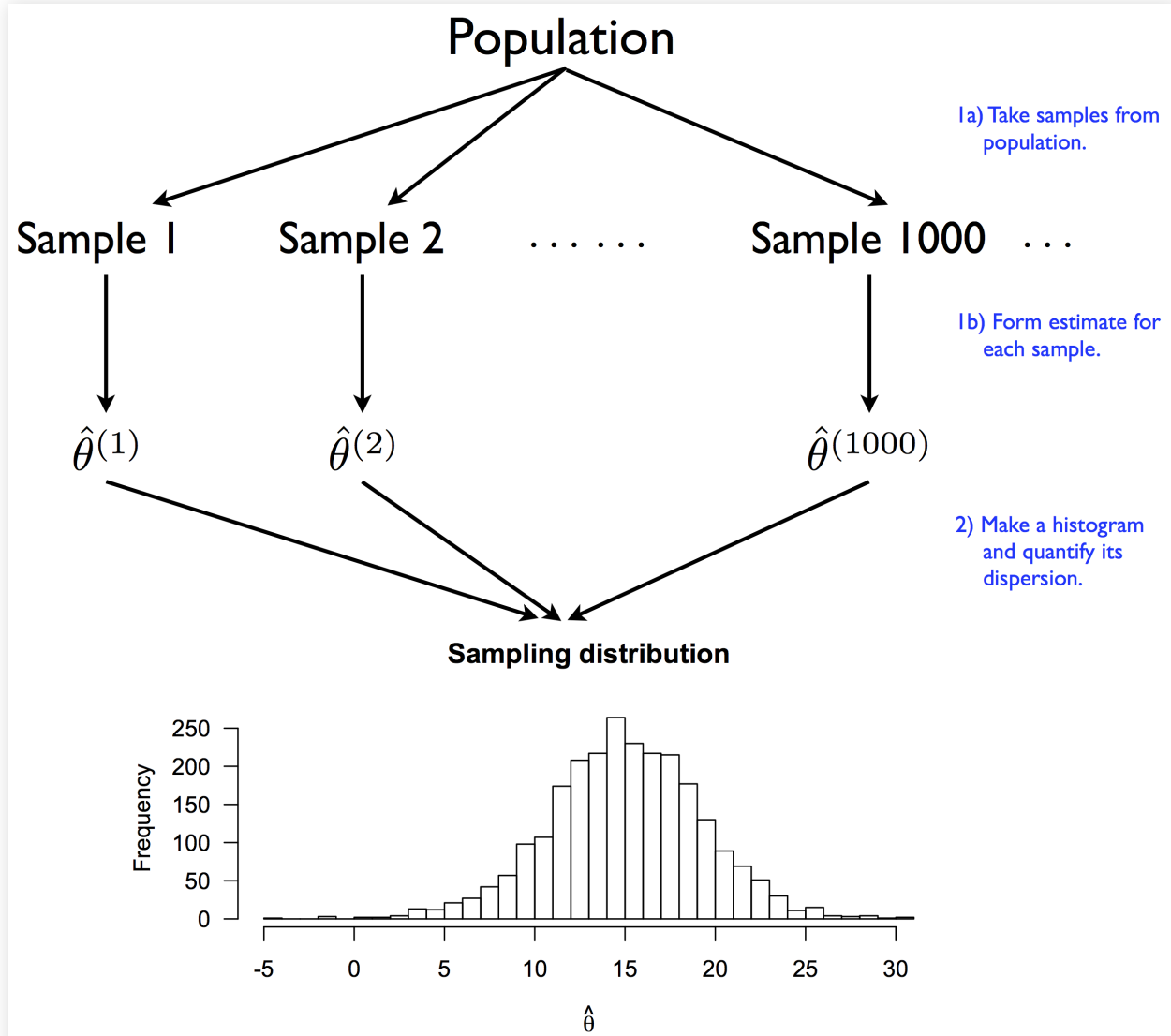


Sampling with replacement is key!

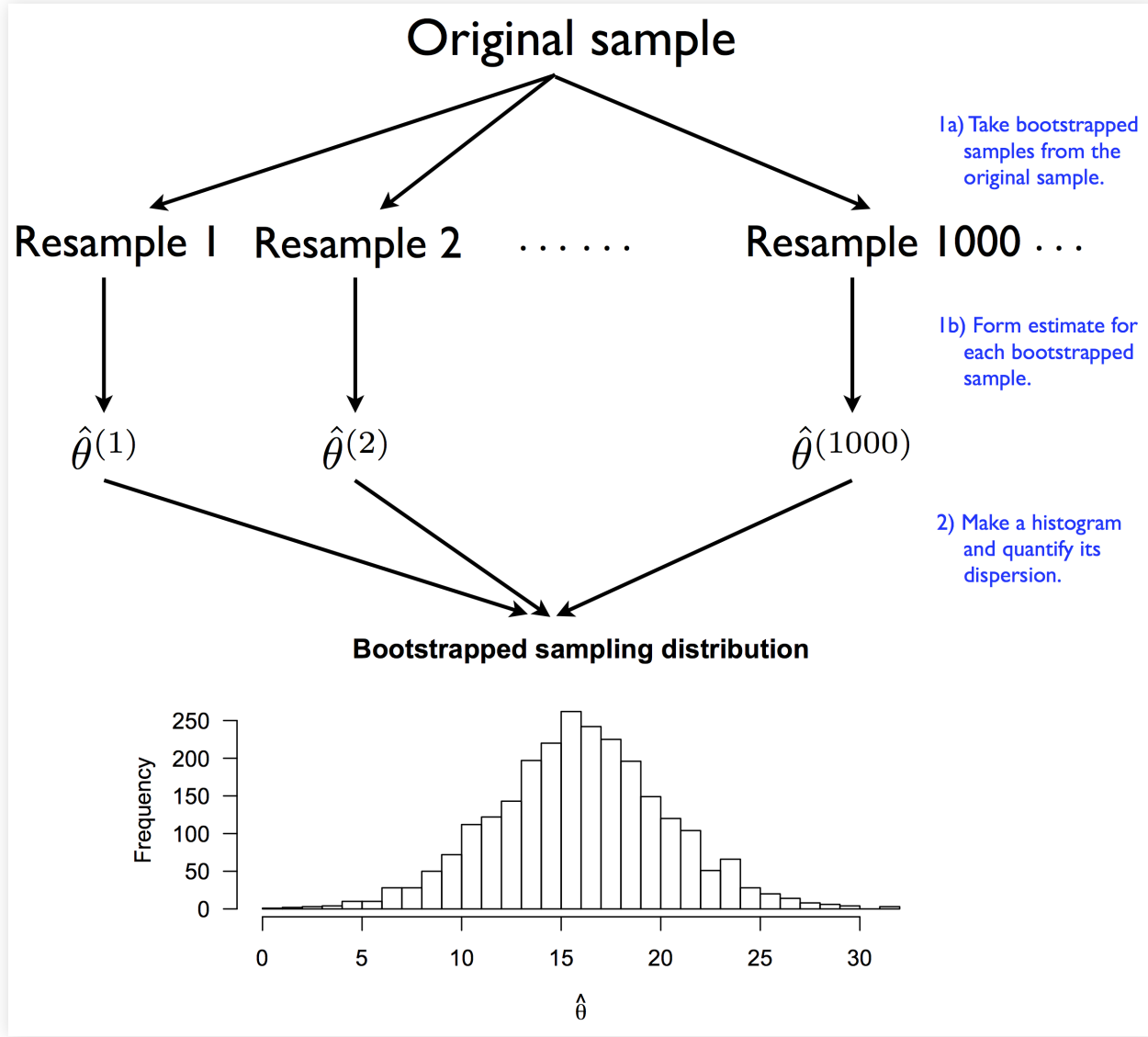
Bootstrapped sample 3



The true sampling distribution



The bootstrapped sampling distribution



The bootstrapped sampling distribution

- Each bootstrapped sample has its own pattern of duplicates and omissions from the original sample.
- These duplicates and omissions create variability in $\hat{\theta}$ from one bootstrapped sample to the next.
- This variability mimics the *true* sampling variability you'd expect to see across real repeated samples from the population.

Bootstrapping: pseudo-code

- Start with your original sample $S = \{X_1, \ldots, X_N\}$ and original estimate $\hat{\theta}_N$.

- For $b = 1, \ldots, B$:

  1. Take a bootstrapped sample $S^{(b)} = \{X_1^{(b)}, \ldots, X_N^{(b)}\}$
  2. Use $S^{(b)}$ to re-form the estimate $\hat{\theta}_N^{(b)}$.

- Result: a set of B different estimates $\hat{\theta}_N^{(1)}, \hat{\theta}_N^{(b)}, \ldots, \hat{\theta}_N^{(B)}$ that approximate the sampling distribution of $\hat{\theta}_N$.

Then what?

Calculate the *bootstrapped standard error* as the standard deviation of the bootstrapped estimates:

$$\hat{se}(\hat{\theta}_N) = \text{std dev}\left(\hat{\theta}_N^{(1)}, \hat{\theta}_N^{(b)}, \ldots, \hat{\theta}_N^{(B)}\right)$$

This isn't the true standard error, but it's often a good approximation!

Example

Let's dig in to some R code and data: `creatinine_bootstrap.R` and `creatinine.csv` (both on class website).

We'll bootstrap two estimators:

- the sample mean
- the OLS estimate of a slope

Confidence intervals

Informally, an interval estimate is a range of plausible values for the parameter of interest. For example:

- Go out some multiple k of the bootstrapped standard error from your estimate:

$$\theta \in \hat{\theta}_N \pm k \cdot \hat{se}(\hat{\theta}_N)$$

- Use the quantiles (e.g. the 2.5 and 97.5 percentiles) of the bootstrapped sampling distribution, to cover a large fraction (e.g. 95%) of the bootstrapped estimates:

$$\theta \in (q_{2.5}, q_{97.5})$$

Confidence intervals

We'd like to be able to associate a *confidence level* with an interval estimate like this. How?

If an interval estimate satisfies the *frequentist coverage principle*, we call it a confidence interval:

> Frequentist coverage principle: If you were to analyze one data set after another for the rest of your life, and you were to quote X% confidence intervals for every estimate you made, those intervals should cover their corresponding true values at least X% of the time. Here X can be any number between 0 and 100.

Confidence intervals

An interval estimate takes the form $\hat{I}_N = [\hat{L}_N, \hat{U}_N]$. Just like a point estimate $\hat{\theta}_N$, the interval estimate is a random variable, because its endpoints are functions of a random sample.

We say that $[\hat{L}_N, \hat{U}_N]$ is a confidence interval at **coverage level** $1 - \alpha$ if, for every θ,

$$P_\theta \left(\theta \in [\hat{L}_N, \hat{U}_N] \right) \geq 1 - \alpha \,,$$

where P_θ is the probability distribution of the data, assuming that the true parameter is equal to θ.

Confidence intervals

The key statement here can be one of the most confusing in all of statistics:

$$P_\theta \left(\theta \in [\hat{L}_N, \hat{U}_N] \right) \geq 1 - \alpha \,,$$

Three questions to ask yourself:

- What is fixed?
- What is random?
- What is the source of this randomness?

Bootstrapped confidence intervals

So recall our two methods of generating an interval estimate using the bootstrap:

- The standard error method: $\theta \in \hat{\theta}_N \pm z^\star \cdot \hat{se}(\hat{\theta}_N)$, where $z^\star$ is a pre-specified quantile of the normal distribution.
- The quantile method (e.g. the 2.5 and 97.5 percentiles of the bootstrapped sampling distribution)

The obvious question is: do these interval estimates satisfy the frequentist coverage principle?

Bootstrapped confidence intervals

The answer is: not always, but often!

In lots of common situations, both forms of bootstrapped interval estimate *approximately* satisfy the coverage requirement:

$$P_\theta\left(\theta \in [\hat{L}_N, \hat{U}_N]\right) \approx 1 - \alpha\,,$$

And the approximation gets better with larger sample sizes. That is, as N gets large,

$$P_\theta\left(\theta \in [\hat{L}_N, \hat{U}_N]\right) \to 1 - \alpha$$

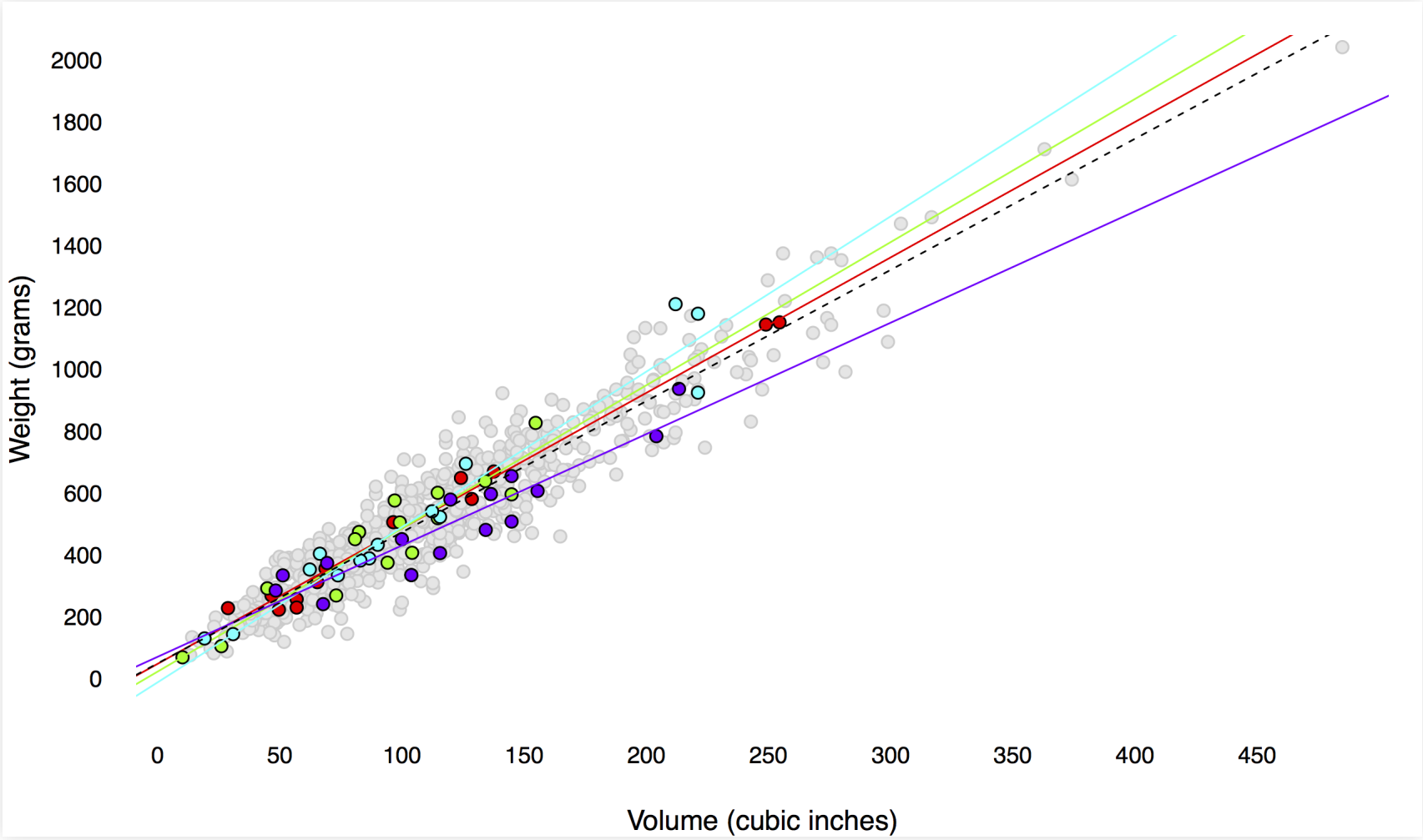
Bootstrapped confidence intervals

The math here is super hairy; we won't go into it. (Google "empirical process theory" if want to learn and you've got a year or two to spare…)

But we can run a sanity check through Monte Carlo simulation!

Let's revisit our thought experiment about fishing…

Bootstrapped confidence intervals

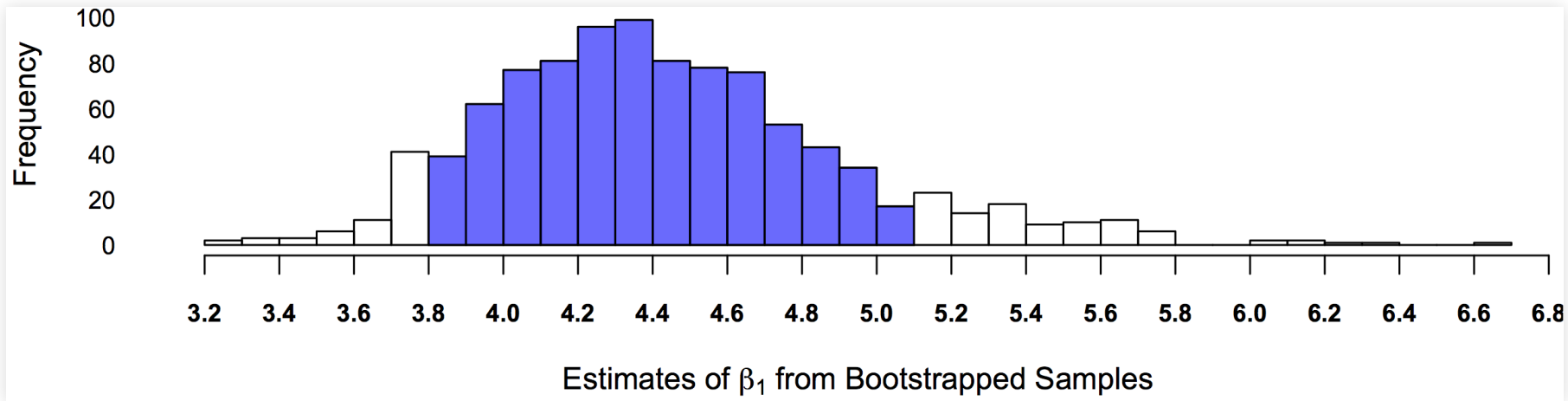


Bootstrapped confidence intervals

Let's go on a 100 fishing trips. On each trip:

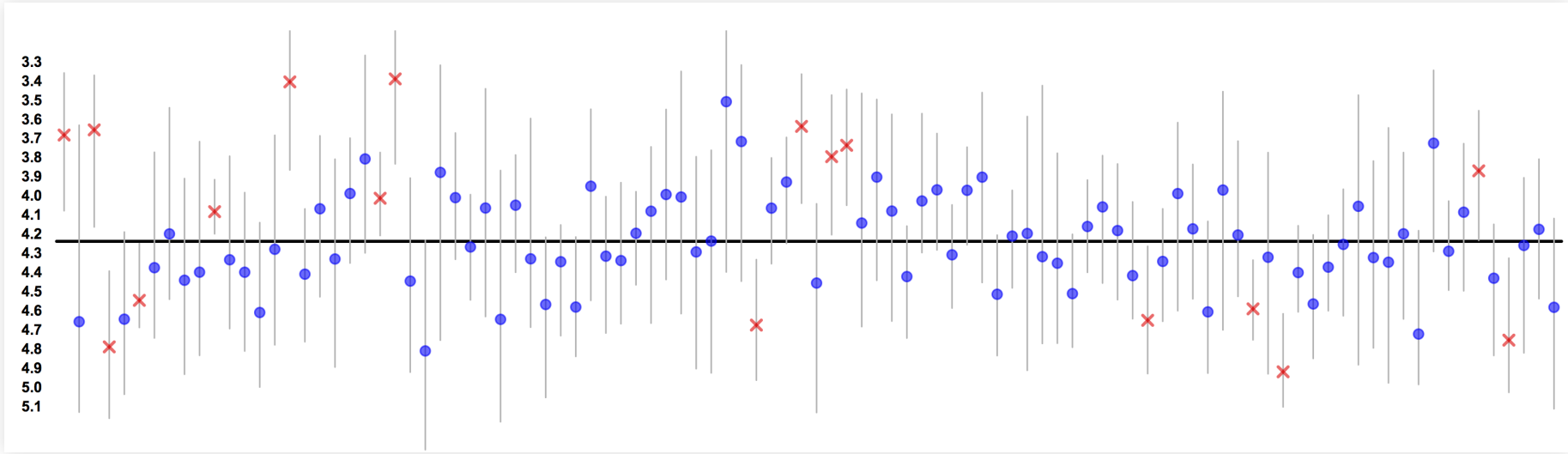
- we catch a sample of $N = 30$ fish
- we run OLS on our sample to estimate β_1: the slope of the weight-vs.-volume line
- we bootstrap our sample to get a 80% confidence interval for the slope.

Bootstrapped confidence intervals



Because we know the slope of the true line ($\beta_1 = 4.25$), we can check whether each bootstrapped confidence interval contains the true value. About 80% of them should!

Bootstrapped confidence intervals



- 100 different samples
- 100 different 80% confidence intervals
- 83 of them cover the truth—pretty good!

Plug-in standard errors

Sometimes we can use probability theory to calculate a “plug-in” estimate of an estimator's standard error. Some simple cases include:

- means and differences of means
- proportions and differences of proportions

Let's see an example and compare the result with a bootstrap estimate of the standard error.

Plug-in standard errors

Suppose that $X_1, X_2, \ldots, X_N$ are a sample of independent, identically distributed (IID) random variables with unknown mean μ and variance σ^2. Let $\bar{X}_N$ be the sample mean:

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Clearly $\bar{X}_N$ is a sensible estimate of μ, since it is unbiased: $E(\bar{X}_N) = \mu$ (show this!)

Plug-in standard errors

We can also calculate the theoretical variance of $\bar{X}_N$ as:

$$\begin{aligned}\text{var}(\bar{X}_N) = \text{var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) &= \frac{1}{N^2}\text{var}\left(\sum_{i=1}^{N} X_i\right) \\ &= \frac{1}{N^2} N\sigma^2 \\ &= \frac{\sigma^2}{N}\end{aligned}$$

Plug-in standard errors

This tells us that the *true standard error* of the sample mean is:

$$\text{se}(\bar{X}_N) = \frac{\sigma}{\sqrt{N}}$$

Or in words:

$$\text{Average error of the sample mean} = \frac{\text{Average error of a single measurement}}{\text{Square root of sample size}}$$

This is sometimes called de Moivre's equation, after Abraham de Moivre.

Plug-in standard errors

There's only one problem with de Moivre's equation: we don't know the true σ!

$$\text{se}(\bar{X}_N) = \frac{\sigma}{\sqrt{N}}$$

The obvious solution is to estimate σ from the data. This results in the so-called "plug-in" estimate of the standard error:

$$\hat{\text{se}}(\bar{X}_N) = \frac{\hat{\sigma}}{\sqrt{N}}$$

where $\hat{\sigma}$ is an estimate of the population standard deviation (e.g. the sample standard deviation).

Plug-in standard errors

Suppose we have an estimator $\hat{\theta}_N$ and we want to know its standard error. The general plug-in procedure involves three steps:

1. Use probability theory to derive an expression for the *true standard error* $\mathrm{se}(\hat{\theta}_N)$.
2. Use the data to estimate any unknown population parameters ϕ that appear in the expression for $\mathrm{se}(\hat{\theta}_N)$. (Note: this might even include θ, the parameter of interest itself.)
3. Plug in the estimate $\hat{\phi}$ into this expression to yield the plug-in standard error, $\hat{se}(\hat{\theta}_N)$

Let's see an example in `predimed_plugin.R`.

Plug-in standard errors

Your turn! For the same `predimed` data set:

- Let p_1 be the true population proportion of people we expect to experience a cardiac event in the control group, and let p_2 be the same proportion in "Mediterranean diet + VOO" group.
- Suppose we're interested in $p_1 - p_2$, the difference in (unknown) true proportions. We estimate this using $\hat{p}_1 - \hat{p}_2$, the difference in *sample* proportions from our data.
- Use the plug-in procedure to derive an estimated standard error, $\hat{se}(\hat{p}_1 - \hat{p}_2)$, and calculate this for the `predimed` data.
- Compare this to a bootstrapped standard error.

Summary

- Any estimator $\hat{\theta}_N$ is a random variable.
- Its probability distribution is called the *sampling distribution*.
- The sampling distribution describes the results of a thought experiment: *what if* we took lots and lots of samples, each of size N, and tracked how much our estimate changed?

Summary

- The *standard error* is the standard deviation of the sampling distribution.
- Roughly speaking, it answers the question: how far off do I expect my estimate to be from the truth?
- A practical way of estimating the standard error is by *bootstrapping*: repeatedly re-sampling with replacement from the original sample, and re-calculating the estimate each time.
- In simple cases we can also calculate a *plug-in* estimate of the standard error, using probability theory together with sample estimates of unknown parameters. **You will do this all the time in Econometrics!**

Summary

- From the bootstrapped sampling distribution, we can get an interval estimate for the parameter of interest (using the standard-error method or the quantile method).
- Both methods approximately satisfy the frequentist coverage principle: under repeated sampling, they contain the true value roughly the correct percentage of the time.
- I tend to use the quantile method because it's pretty intuitive!

Bonus: parametric bootstrap

The essential idea of the bootstrap is to simulate synthetic data sets by resampling from the original sample. This is often called the *nonparametric bootstrap*.

A common variation is called the *parametric bootstrap*: simulate synthetic data sets by simulating from a fitted parametric model.

See `predimed_bootstrap.R`!